DOCUMENT RESUME

ED 445 012 TM 031 549

AUTHOR Impara, James C.; Plake, Barbara S.
TITLE A Comparison of Cut Scores Using Multiple Standard Setting Methods.
PUB DATE 2000-04-0
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (81st, New Orleans, LA, April 24-2, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Intermediate Grades; *Mathematics Tests; Scoring Formulas; Standards
IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

This paper reports the results of using several alternative methods of setting cut scores. The methods used were: (1) a variation of the Angoff method (1971); (2) a variation of the borderline group method; and (3) an advanced impact method (G. Gillon, 1996). The results discussed are from studies undertaken to set the cut scores for fourth grade reading and mathematics in a public school system. Twenty-two elementary school teachers served as judges. The three methods tended to result in similar cut scores with certain predictable variations. When the cut score was set in the lower tail of the distribution, the Angoff method tended to result in a cut score that was lower than the cut score set using the borderline group method. Recommendations are made for supplementing the Angoff method with additional data from alternative methods to improve the appropriateness of this method when setting performance standards in school settings. (Contains 15 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 012

J. Impara

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Comparison of Cut Scores using Multiple Standard Setting Methods[1]

James C. Impara
University of Nebraska - Lincoln

Barbara S. Plake
University of Nebraska - Lincoln

Author Contact:
James C. Impara
Buros Institute for Assessment Consultation and Outreach
135 Bancroft Hall
University of Nebraska - Lincoln
Lincoln, NE 68588-0353
(402) 472-8804
email: jimpara@unl.edu

TM031549

BEST COPY AVAILABLE

Abstract

This paper reports the results of using several alternative methods of setting cut scores. The methods used are a) a variation of the Angoff (1971) method, b) a variation of the borderline group method, and c) an advanced impact method (Dillon, 1996). The three methods tend to result in similar cut scores with certain predictable variations. When the cut score is set in the lower tail of the distribution, the Angoff method tends to result in a cut score that is lower than the cut score set using the borderline group method. Recommendations are made for supplementing the Angoff method with additional data from alternative methods to improve the appropriateness of this method when setting performance standards in school settings.

A Comparison of Cut Scores using Multiple Standard Setting Methods

School districts are experiencing increased pressure to use results from assessment programs to identify students who do not have the needed skills to graduate from high school or who may have problems in "the next grade" and may benefit from instructional activities beyond those provided in their regular classroom. These policies are often based, in part, on student test performance. Students with scores lower than the minimum passing score (MPS) are classified as needing instructional interventions beyond what the regular classroom teacher can provide (e.g., interventions like summer school, or after school programs). These MPS are often determined by using the Angoff standard setting method.

The Angoff method may be chosen because it is considered to be defensible, easy to use, easy to explain to the policy makers who may ultimately set the passing score (Berk, 1986), and it has been found to be extremely replicable across panels (Mehrens, 1995). Shepard (1995), however, suggests that a long history and ease of use do not justify continued use of a method that may produce cut scores that result in too many invalid decisions. She raises this point in the context of studies done to investigate the appropriateness of using the Angoff standard setting method in the National Assessment of Educational Progress (NAEP). In addition to Shepard's concern, the Angoff method can be expensive for a school district due to the requirement that teachers who will serve as "judges" need to meet together for training and for the standard setting process. This meeting may require extra pay for teachers or paying for substitutes while the teachers are performing the standard setting activities.

This paper compares the results of using the Angoff method of setting a cut score with two other methods. First, the various methods are described and the theoretical expectations for the resulting cut scores are explained. Then the results of the application of these methods are shown and discussed. Finally, suggestions for combining multiple methods to provide a rational range of cut score values within which the final standard may be set are proposed.

Overview of the Angoff method

In its basic form, the Angoff (1971) method of setting cut scores entails asking a group of judges to examine each item on a test and to estimate what proportion of a target group of examinees will answer each item correctly. The target group of examinees are those who are on the borderline between the "competent" and the "incompetent." Often judges are instructed to envision a hypothetical group of 100 target examinees and directed to estimate how many of this hypothetical group will answer each item correctly. These item performance estimates for the target group are summed across items to obtain each judge's cut score. The judges' cut scores are averaged to obtain the estimated minimum passing score (MPS) that a minimally competent candidate (MCC) would obtain. In a school setting the language "minimally competent candidate" has little meaning. Instead some school districts substitute phrases such as "just competent student" or "barely proficient student."

Extensive research on the Angoff method has resulted in numerous modifications. Among the most frequent modifications are: a) providing extensive training for judges in the process of both identifying the target examinees and in estimating their item performance, b) providing actual performance data to the judges (often along with the impact -- percent passing or failing -- associated with the judges' initial cut score), and c) including more than one opportunity to estimate examinee performance (Plake, 1998). Other modifications that are less pervasive include permitting judges to discuss their ratings or perspectives after an initial round of item performance estimation, and requiring judges to estimate performance for a category of examinees in addition to the target examinees (e.g., estimating performance for the average examinee in addition to the target examinee). The studies reported in this paper used the variation described in Impara and Plake (1997) in which judges made dichotomous estimates of examinee performance.

Recent research suggests that in some standard setting contexts the standard setting judges make item performance estimates that systematically differ from actual performance. Shepard (1995) reports that judges' ratings are

less extreme than actual performance. For example, if 90% of the examinees actually answer correctly (an easy item) judges will tend to estimate that a lower percent will answer correctly (e.g., 85%). At the other extreme, items that are difficult (e.g., only 30% answer correctly) judge estimate that the item will be easier (e.g., 35% may be estimated to answer correctly). The impact of this phenomenon would be that, other things being equal, too many examinees would pass if the test consisted mostly of hard items and too many would fail if the test were comprised of easy items[1].

Linn and Shepard (1997) undertook a simulation study in an attempt to understand Shepard's findings that judges' cut scores were not consistent with their expectations for the impact of the cut scores. Based on their simulation, Linn and Shepard (1997) report that Angoff ratings for examinees whose estimated score is below the mean (e.g., in NAEP the below basic students) will systematically be lower than expected (and vice versa for the cut score used to identify students in the upper tail of the distribution). The extent of the difference between judges' expectations and the actual impact of the cut score is a function of the judges' expectations, the intercorrelation of items (the lower the inter-item correlations the greater the disparity), and the length of the test (the longer the test, the greater the disparity).

From the simulations, it is reasonable to predict that when the target examinees are expected to score below the mean the cut score set by the Angoff method will be set too low (too many will pass). Conversely, for target examinees, who are expected to score above the mean, the cut score will be set too high (too few will pass). In most school settings, the cut score is generally located in the lower tail of the distribution when focusing on graduation, promotion, or identification of students needing additional instructional exposure. This leads to the concern that Angoff based cut scores may be set too low.

Alternatives to the Angoff method

There are several other methods of setting cut scores that might be used in a school setting. Only those methods that are widely accepted and that could be accomplished at a cost that would not exceed the cost of the Angoff method are considered here. Specifically, borderline group method (Livingston and Zieky, 1982), and the advanced impact method (Dillon, 1996). Each of these methods is described very briefly and its use in a school setting either independently or as a complement to the Angoff method is discussed.

Borderline Group method (Livingston and Zieky, 1982)

This method of setting a cut score can be accomplished in several ways. The method described below is a modification of the method described in Livingston and Zieky.

In a school setting, teachers may be provided with a list of the students in their classes who will take the test on which a cut score is to be set. After providing teachers with a description of the test content, teachers are directed to make global estimates of their students' performance on the test. These global estimates might classify students into categories such as "below proficient", "proficient", and "beyond proficient". Each of these categories would be defined operationally within the context of the test content (the definitions may be drawn up by a committee of teachers or by central office staff). After making the initial classifications, the teachers are asked to go back through the list and (for example) indicate which proficient and below proficient students are on the borderline between these two categories. For the borderline group method students who are identified in this final classification comprise the borderline group. Test performance of students in this groups serves as the basic data for setting the cut score. This is not the only way to identify the students who are in the borderline group, but it is a strategy that has been shown to work in a school setting (Crawford and Spangler, 1997).

All classifications must be completed prior to teachers knowing the scores of their students on the test. It may be done prior to testing, or it may be done in conjunction with the Angoff method at the time item performance estimates are made. An advantage of this method is that it is a task that is consistent with what the literature (e.g., Hoge and Coladarci, 1989) indicates teachers can perform

successfully. If used in conjunction with the Angoff method the teachers will have a common conceptualization of the target examinees being used for each method[2].

Although research has shown that different standard setting methods produce different results, some systematic differences have been noted. Livingston and Zieky (1989) and Jaeger (1989) compared the results of different methods and found, for example, that the borderline group method typically resulted in a cut score that was higher than the cut score obtained from applying the Angoff method.

One reason for these systematic differences is suggested by examining both Livingston's (1995) study of the borderline group method and Linn and Shepard's (1997) analysis of the Angoff method. Livingston (1995) suggests the borderline group method is susceptible to regression to the mean. He argues that regression to the mean will occur when students have been preselected by teachers based on their prior low achievement and subsequently take an achievement test. The mean on the achievement test is regressed toward the mean of the total group thus resulting in a cut score that would be "too high" when applied to future groups. Livingston's observations combined with the research described above by Linn and Shepard (1997) suggests that the borderline group method will tend to result in a systematically higher cut score than the Angoff method (when the cut score is in the lower tail of the distribution). Because the cut score from the borderline group method, due to regression to the mean, will be too high, the borderline group method may be considered an upper bound for the cut score and the cut score derived from the Angoff method could be considered a lower bound. For cut scores needed to identify examinees at the upper end of the distribution, the relative positions of the two cut scores from these methods would be reversed (i.e., the borderline group method will result in a lower bound and the Angoff method will result in an upper bound).

Advanced Impact Method (Dillon, 1996[3])

Another method that might be used to set the cut score is to simply ask teachers what percentage of their current students are ready to graduate, or are

eligible for promotion to the next grade, or who qualify for extra instructional assistance. Prior to asking teachers to estimate the percentage of their students who "qualify," it is important to define what it means to qualify so that all teachers are using the same basis for their estimates. The question could be asked at the same time as the teachers are classifying their students into the global categories to be used for the borderline group method. If this is done at the time the Angoff ratings are collected, it can be done either just following the training, but prior to round 1 of the Angoff item performance estimates, or it could be done just prior to providing actual performance data prior to round 2 of the Angoff ratings.

Averaging across teachers will result in an estimate of the percentage of students in the district who are eligible for instructional intervention. After administering the test a cumulative percentage distribution can be used to find the score that identifies the appropriate percentage of students. One might expect the cut score obtained by this method to be near the cut scores set by either the Angoff method or the borderline group method. To the extent that this estimate is more extreme (either higher or lower) it may reflect a boundary point. There is a risk that this method may result in some deflation of the appropriate value if there is some belief by the teachers that the percentage of qualified students will reflect badly on them. Similarly, if the teachers define the students in need in a different way than the district intends, then the percent of students in need may be highly inflated (almost any student not grasping all the concepts in the content area may be classified as being in need of extra instruction). For these reasons, this method should not be the only method employed; it should only be used as a supplement to other methods and extreme values may need to be discounted.

The empirical studies

The combination of the three methods described above for setting a cut score were used in the Millard schools in several cut score studies. The results from two of these studies are reported below. These results are reflective of the results that have been observed in every case (including several studies done in

school systems other than Millard). The only variations are that on some occasions the Advanced Impact method has been more slightly more extreme than either the Angoff or the Borderline Group method. One of these cases is illustrated in the data shown.

The results discussed are from studies undertaken to set the cut score in 4th grade reading and in 4th grade mathematics in Millard Public Schools. Both studies were conducted in spring, 1999. These two studies were two in a long series of cut score studies and other collaborations between this school system and the Buros Center for Testing. Only the procedures used in the mathematics study are described in detail because both studies followed essentially the same procedures.

Procedures

The Mathematics Test

The mathematics test was developed by the school system. The test consists of approximately 60 multiple choice items. It is designed to assess a set of learning outcomes defined by the school district. The test was developed and pilot tested within the district. The psychometric characteristics of the test were of sufficient quality to justify its administration and use as one element in the identification of students who might be eligible for instructional interventions beyond what would normally be available to the regular classroom teacher. These interventions might include recommendations for summer school, specially designed after school programs or other activities (other than a special education classification) to try to bring the student's performance up to standard. The use of a test (and cut score) to make decisions about students who needed to be "relooped" has been a fixture in the school district for several years.

Teachers

There were 22 elementary teachers selected from among the school district's fourth grade teaching staff[4]. Among these 22 teachers, some had participated in standard setting studies in past years, but most had not had this experience previously.

Data collection

The standard setting studies (called workshops) were held in the school districts' administrative building. The studies began promptly at 8:00 a.m. and lasted until after lunch. The initial activities involved orienting the teachers to the purpose of the workshop and the activities they would be participating in during the day. The orientation involved providing the teachers with the district's general definition of the target student [the "Barely Proficient Student" (BPS)]. This was followed by providing the teachers with a review of the table of specifications of the test. The teachers then engaged in an extended discussion of the characteristics of the BPS within the context of the table of specifications.

Teachers were asked to think about a student in their class who was a BPS. They were to keep this student in mind as they made item performance estimates on both a practice test and the operational test. This process was followed by a practice exercise that permitted the teachers to experience the process on 6 to 9 test items that had been administered previously in the district, but were not part of the operational test. These practice items were selected to reflect the range of difficulty that would be found in the operational test. After making their round 1 estimates of whether the BPS would answer each question correctly or not (consistent with Impara and Plake, 1997), teachers shared their estimates with the group. The actual proportion of the students in the district was shown (the total group p-value) along with the teachers' average estimated p-value for the BPS. There was a discussion about each item why the item was hard or easy for the BPS in the context of the earlier discussion about the characteristics of the BPS and the test's table of specifications. After all practice items were discussed, a cut score for the practice test was computed and displayed on a cumulative frequency distribution of the scores on the practice test. This distribution was explained to the teachers in terms of the percent of the districts' students who would pass if this were the operational test and if there were no more opportunities to change the cut score. They were then provided an opportunity to make a second round of item performance estimates.

Once the practice was completed, the next step was for the teachers to make their estimates for the Advanced Impact Method. Teachers were provided a form and asked to estimate for both their class and the district as a whole, the

percentage of students who they believed would be classified as needing instructional interventions.

When the teachers had completed this form, they were provided copies of the operational tests and forms on which to write their item performance estimates for the Angoff method. They completed their round 1 estimates and these estimates were entered into an Excel spreadsheet prepared in advance for this study. After all round 1 estimates were entered and a cut score computed, these results as well as item p-values for district students were provided and explained. After questions about the data were answered, the teachers made their round two item performance estimates without discussion.

After their round 2 Angoff estimates were made, teachers were provided their class roles and were asked to classify their students in terms of the four performance categories described above in the discussion of the Borderline Group Method (below proficient, barely proficient, proficient, and beyond proficient). When this task was finished, teachers completed a form used to evaluate the workshop. This evaluation asked about their perceptions of the adequacy of the training, their comfort level with the process, and their confidence in the cut score that would result from the process. Teachers were then dismissed to return to their buildings.

Results

The results, as shown in Table 1, are consistent with the expectations described in the introduction. That is, for both the reading and the mathematics tests, the cut score resulting from the Angoff method is lower than the cut score resulting from the Borderline Group method.

The teachers who participated in the standard setting workshop for reading estimated that slightly more students in the district would be Below Proficient than they estimated for their collective classes[5]. Transforming these percentages into a passing score from the cumulative frequency distribution resulted in the same cut score. The Borderline Group cut score in reading was slightly higher than the cut score that resulted from the Angoff method. For the mathematics test, the cut score from the Advanced Impact Method for teachers'

classes was slightly lower than the Angoff estimate and that cut score estimated for students in the district as a whole was slightly above the Angoff cut score.

The values obtained from these three standard setting methods are all reasonable. Because they are reasonable, they provide a range of scores within which the policy decision associated with setting the final cut score can be made.

Table 1. Cut scores set by different methods in Reading and Mathematics.

	Reading		Mathematics	
Method	Cut score	Percent of students failing	Cut score	Percent of students failing
Initial Impact (class)	23	13.0%	40	19.6%
Initial Impact (dist.)	23	13.8%	42	21.5%
Angoff	22	9.2%	41.5	20.8%
Borderline Group	27	24.9%	48	30.4%

Conclusions and Recommendations

Three methods were used in these two studies to estimate the test score that could be used to identify students who could benefit from special instructional interventions. It was expected that the three methods would result in slightly different cut scores. Specifically, based on prior research, the cut score from the Borderline Group method was expected to be higher than the cut score from the Angoff method. A third method, the Advanced Estimate method, has not been reported in the literature in studies that compared it with other methods. Our findings were consistent with expectations related to the relative positions of the Borderline Group and the Angoff methods. We observed that the Advanced Estimate method also produced cut scores at or near the cut scores from the Angoff method. Moreover, we observed that teachers tended to have a halo effect regarding the estimates of the percentage of students in their classes who would be classified by the test as being Below Proficient when compared to the percentage of students in the district who would be classified as Below Proficient.

Our results reinforce the recommendations made by others to use multiple methods (see for example, Jaeger, 1989). More recently, Livingston (1995) advocates using both the borderline group and contrasting groups methods in order to minimize the bias he describes when only the borderline group method is used. In a school setting, borderline or contrasting groups methods may be very reasonable approaches especially in subject areas where additional complementary data might be collected and reported that increase the confidence levels in the cut scores. Such collateral data might be student grades (the percentage of students receiving grades consistent with the decisions that would be made using the cut score), student performance on different tests, and scores of students who are in classes that would represent differential levels of performance. For example, in high school mathematics there is a continuum of courses that represent increasing levels of difficulty and competence in mathematics, thus one might look at the rank order of average scores obtained by students in the different courses. See Giraud, Impara, Plake, Hertzog, and Spies, 1997 for an illustration of this strategy.

There should be some means of triangulation employed to provide policy makers (who actually set the standard) with a defensible range of values within which to set the cut score. That is, policy makers should be provided more than just a single point estimate or range of values that results from the use of a single standard setting study.

As school systems are being pressured to raise standards and to be accountable for their students' learning, they are turning to the use of assessment programs that may include using a test score to help make critical decisions about the future instructional experiences their students will encounter. Because of the high stakes nature of the decision, the test score used in this decision-making process should not be set capriciously. The Angoff method is a long used and respected method of setting a cut score, but it has recently come under attack because it may result in too many invalid decisions. This paper has provided evidence that multiple methods for determining that cut score can provide reasonable boundaries within which a cut score may be set.

Endnotes

[1] However, in the NAEP, the method of converting the estimates from raw performance estimates to scaled scores that adjusted for difficulty had the opposite impact. That is, the cut score defining the advanced students was "too high" (fewer students than judges expected were classified as advanced) and "too low" for the students classified as basic (although this was less severe of a problem). The too high and too low are based on the percent of examinees judges expected to be classified as advanced or below basic, respectively.

[2] An advantage of this method of obtaining classifications is that it permits some validation of the teacher's ability to make the classifications. That is, if the students who are classified as being below proficient systematically obtain very high scores on the test, the teacher's ratings may be discounted. Some care in making this decision is needed, however, because some students are much more capable than they appear (explaining higher than expected scores) and some more capable students may just "blow off" the test. The overall average scores for students in each classification, however, should rank order such that below proficient student's average is the lowest and beyond proficient students is the highest.

[3] This method is attributed it to Dillon (1996), but he suggests that is it simply a variation on methods used to "correct" the cut score obtained when using the Angoff method. He does not call it the Advance Impact Method.

[4] There were 22 teachers in both the mathematics and reading standard setting studies. These were not the same teachers, but there was some overlap in the sample of teachers. The characteristics of the two samples in terms of experience, representiveness of the district's schools, and other factors were similar.

[5] In this and in other studies we have consistently observed that teachers' estimates of the percentage of their own students who are below proficient (or whatever similar language is used) tend to be lower than their estimates for the district as a whole. We think this is a halo effect suggesting that "My classes are okay, but other teachers are not doing as good a job as I am."

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, p. 508-600. Washington, DC: American Council on Education.

Berk, R. (1986). A consumer's guide to setting performance standards of criterion-referenced tests. Review of Educational Research, 56, 137-172.

Crawford, J. C., & Spangler, S. (1997, March). One school district's experience in setting standards to articulate the measurement of essential learning outcomes. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Dillon, G. F. (1996). The expectations of standard setting judges. CLEAR Exam Review, Summer, 22-26.

Giraud, G., Impara, J. C., Plake, B. S., Hertzog, M., & Spies, R. (1997, October). Cutscore Validity in a Public School Setting. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. Review of Educational Research, 59, 297-313.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34, 355-368.

Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 485-514). Washington, DC: American Council on Education.

Linn, R. L., & Shepard. L. (1997, June). Item-by item standard setting: Misinterpretations of judge's intentions due to less than perfect item inter-correlations. Paper presented at the Large Scale Assessment Conference, Colorado Springs, CO.

Livingston, S. A. (1995). Standards for reporting the educational achievement of groups. Joint Conference on Standard Setting for Large Scale Assessments, Proceedings Volume II. Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1989) A comparative study of standard setting methods . Applied Measurement in Education, 2, 121-141.

Mehrens, W. (1995). Methodological Issues in Standard Setting for Educational Exams. Joint Conference on Standard Setting for Large Scale Assessments, Proceedings Volume II. Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. Applied Measurement in Education, 11(1), 65-80.

Shepard, L. A. (1995). Implications for Standard Setting of the NAE Evaluation of NAEP Achievement Levels. Joint Conference on Standard Setting for Large Scale Assessments, Proceedings Volume II. Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.